BASED ON WSC 2015 TUTORIAL

BOOTSTRAP CONFIDENCE BANDS
AND
GOODNESS-OF-FIT TESTS
IN
SIMULATION INPUT/OUTPUT MODELLING

Russell Cheng

UNIVERSITY OF Southampton

Introduction to Bootstrapping

First let me emphasise that Random Variables are a very important part of the Statistical Uncertainty that occurs in Simulation Modelling. Bootstrapping is a very simple method of studying this uncertainty, which it does by answering a question that we can ask of all Random Variables. Indeed it is the ONLY single question you can ask about a Random Variable.

Bootstrapping is all about answering this single question in a simple way!

Before we look at this question, let me summarise the Course so far: You will have been introduced to the idea of using Statistical Simulation Models to represent Real Life systems of interest. Use of the word Statistical highlights the fact that these systems are subject to statistical variation which can occur in the Input quantities, or which can occur within the System themselves. In a simulation model these quantities are treated as random variables which are generated by random variate generators.

A simple example is an M/M/1 queue where the first M denotes customers who arrive randomly with interarrival times that are exponentially distributed with PDF

$$\lambda e^{-\lambda t}, \ t \geq 0$$

where λ is the mean arrival rate. These interarrival times are the Inputs.

In the M/M/1 queue the customers are served by a single server indicated by the '1'. The times to serve each customer are system generated random quantities with the second M indicating that the service times are also exponentially distributed

$$\mu e^{-\mu t},\ t \geq 0$$

with μ the mean service rate of the server when busy.

In general, the quantity of interest is regarded as Output. This is will depend on the random input quantities and the system generated random quantities, so will also be a quantity that varies statistically.

In our example we might take the quantity of interest to be the mean waiting time in the queue, which happens to be known, with

$$W(\lambda, \mu) = \frac{\lambda}{\mu(\mu - \lambda)}$$

For more complicated queues the formula is not always so simple, which is why the simulation model is needed to estimate the Output value numerically.

Even if the formula for the output is known, numerical estimation is still required to estimate the parameters, as in our M/M/1 queue. We could consider this numerical estimation to be part of Input Modelling, as parameter values are needed as inputs in order to run the Simulation

Model. But our real interest is in estimating the Output and its statistical variability. So estimating the parameters could equally be thought of as part of the Output Modelling.

Personally I think that trying to make a distinction between Input and Output Modelling is unhelpful and confusing. So though I will use these terms I will prefer to focus on Random Variables.

You will have been shown, in the previous lectures, how to estimate parameters using Maximum Likelihood (ML) estimation. This does this by fitting parameters to data. The data can have been obtained in different ways, but will depend on the parameters so are random variables which depend on the parameters. ML works by fitting the probability distributions to the data. Moreover, you will have been shown the attractive property of ML estimators in allowing the accuracy of the ML estimates to be assessed using Asymptotic Normal Theory, which shows that, as more data are obtained the parameter estimates become increasingly close to normally distributed which moreover can be estimated, so that confidence intervals can be obtained that allow one to gauge how accurate are the results.

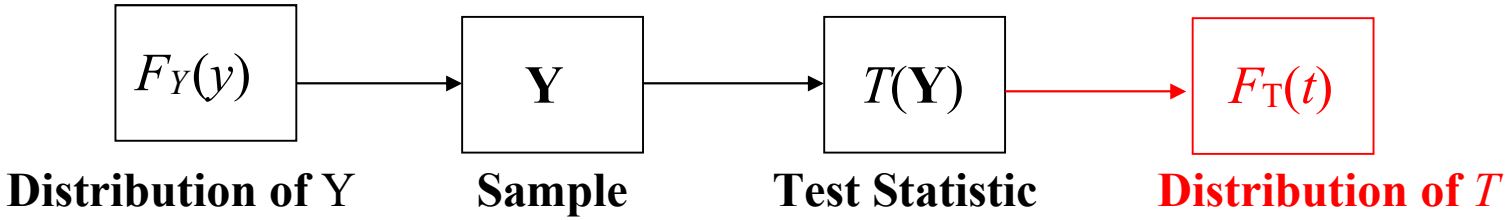
Bootstrapping steps in here as it offers a simple alternative without having to invoke the more complicated mathematics of asymptotic theory

One thing to realise at the outset is that there often is a common misconception that bootstrapping gives you something for nothing and that it somehow allows one to estimate parameters more accurately without having to obtain more data. This has led to an initial mistrust, when bootstrapping was first proposed

Bootstrapping is summarised in Chapter 4 of my book Cheng (2017).

A Point to Note: Though I have introduced bootstrapping as an attractive alternative to asymptotic normal theory when using ML, it has more **general uses**, as it solves the following

Basic Statistical Question



What is the Distribution of $T(\mathbf{Y})$?

Example. Voting in an Election. We have a constituency of voters.
Distribution of interest is the distribution of the votes.
Sample is an Opinion Poll.
Test Statistic of interest to a candidate is
Distribution of Interest is the proportion voting for her/him.

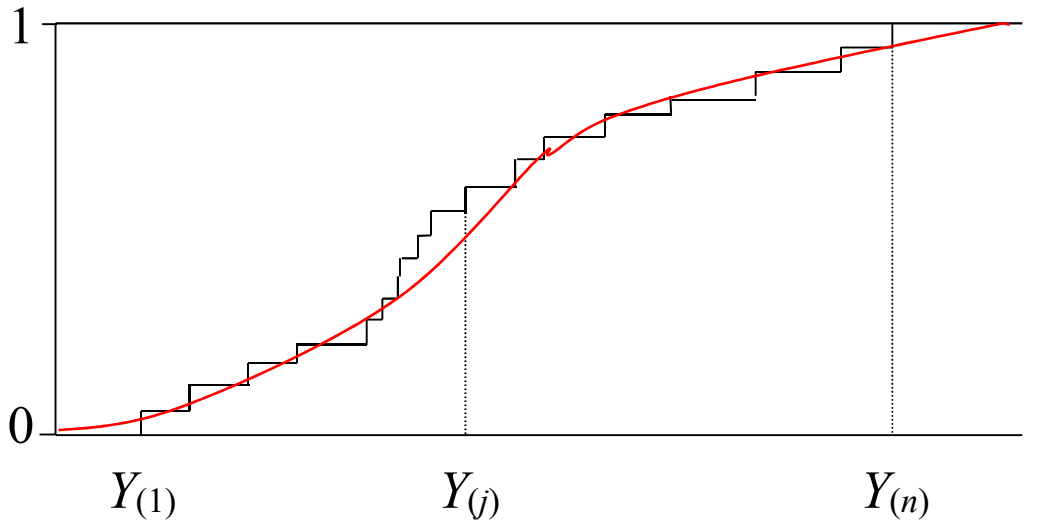
Bootstrapping depends on the properties of:

The ***empirical distribution function*** **(EDF)** defined as

$$\widetilde{F}_Y(y) = \frac{\#\ of\ \ Y's \le y}{n}$$

where $Y_i,\ i = 1, 2, ..., n$ is a *random sample*

EDF of the Y_i, $\widetilde{F}_Y(y)$



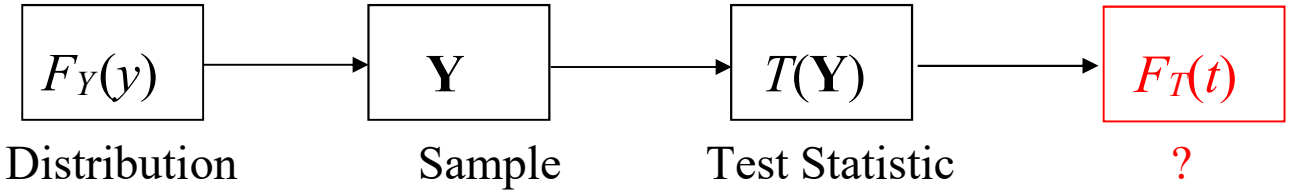
The EDF estimates the *cumulative distribution function* (CDF) of *Y*.

Fundamental Theorem of Sampling

EDF → CDF with probability one, as n→ ∞

? How does this and bootstrapping help with:

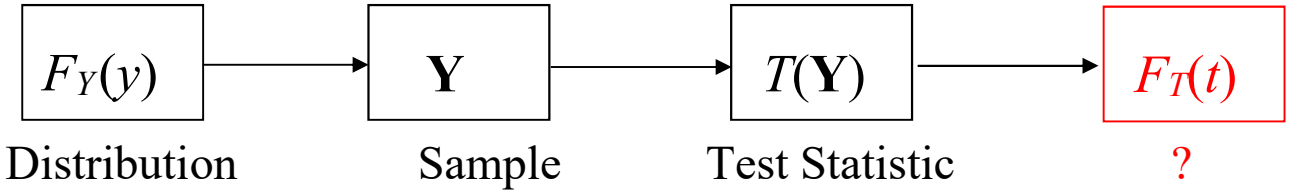
The Basic Sampling Question



What is the Distribution of $T(\mathbf{Y})$?

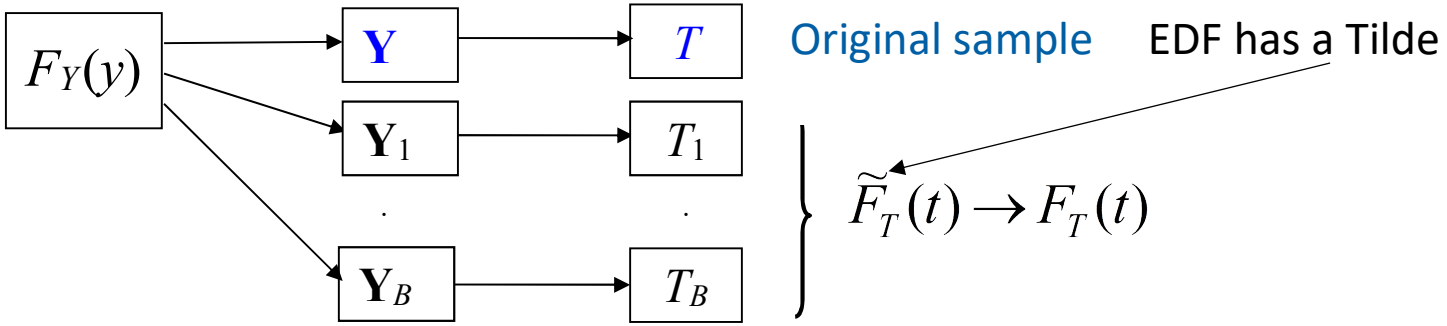
? How does this and bootstrapping help with:

The Basic Sampling Question



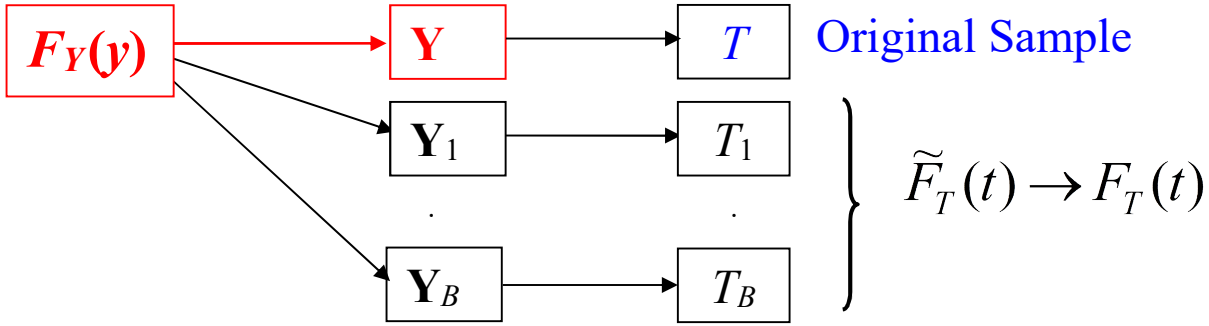
What is the Distribution of $T(\mathbf{Y})$?

The Basic Statistical Question is answered if we could replicate the process a large number of times



Problem: Sampling from the Distribution often difficult (Expensive, time consuming)

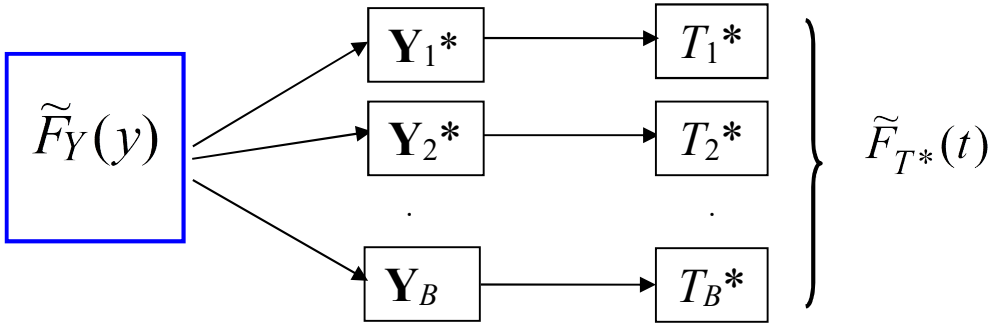
Let us focus on the difficult part:



Note that the Fundamental Theorem applies to this original sample Y:

$$\text{EDF } \widetilde{F}_Y(y) \to F_Y(y) \text{ as } n \to \infty$$

Replace $F_Y(y)$ by EDF $\widetilde{F}_Y(y)$ of original sample to get the Bootstrap Version



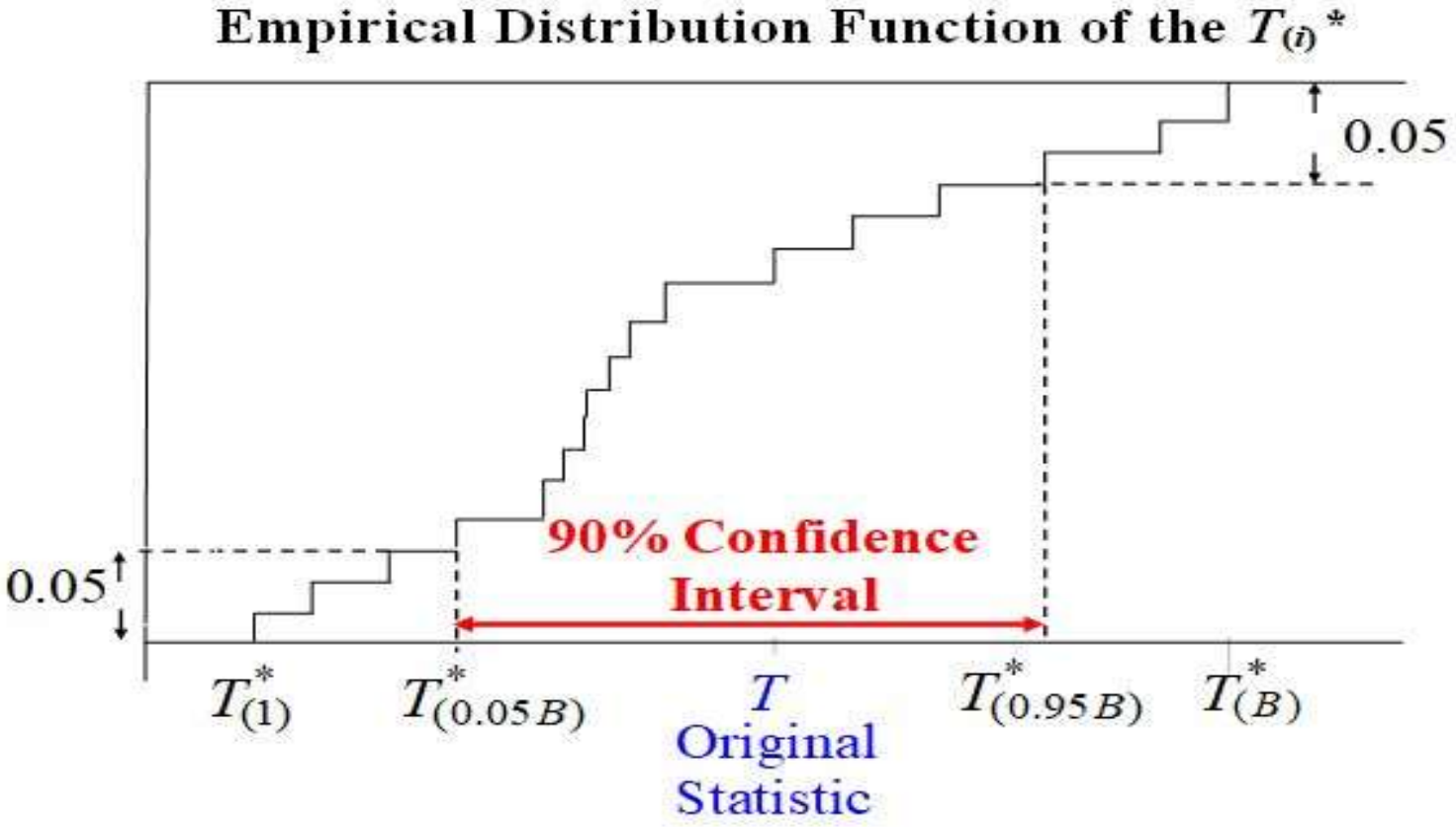
The pseudocode for the entire bootstrap process is as follows:

// $\boldsymbol{y} = (y(1), y(2), ..., y(n))$ is the original sample.
// $T=T(\boldsymbol{y})$ is the calculation that produced T from $\boldsymbol{y}$.

```
For k = 1 to B
{
  For i = 1 to n
  {
    j = Int [1 + n × Unif()]  // Unif ~ U(0,1)
    y*(i) = y(j)
  }
  T*(k) = T(y*)
}
```

Confidence Intervals

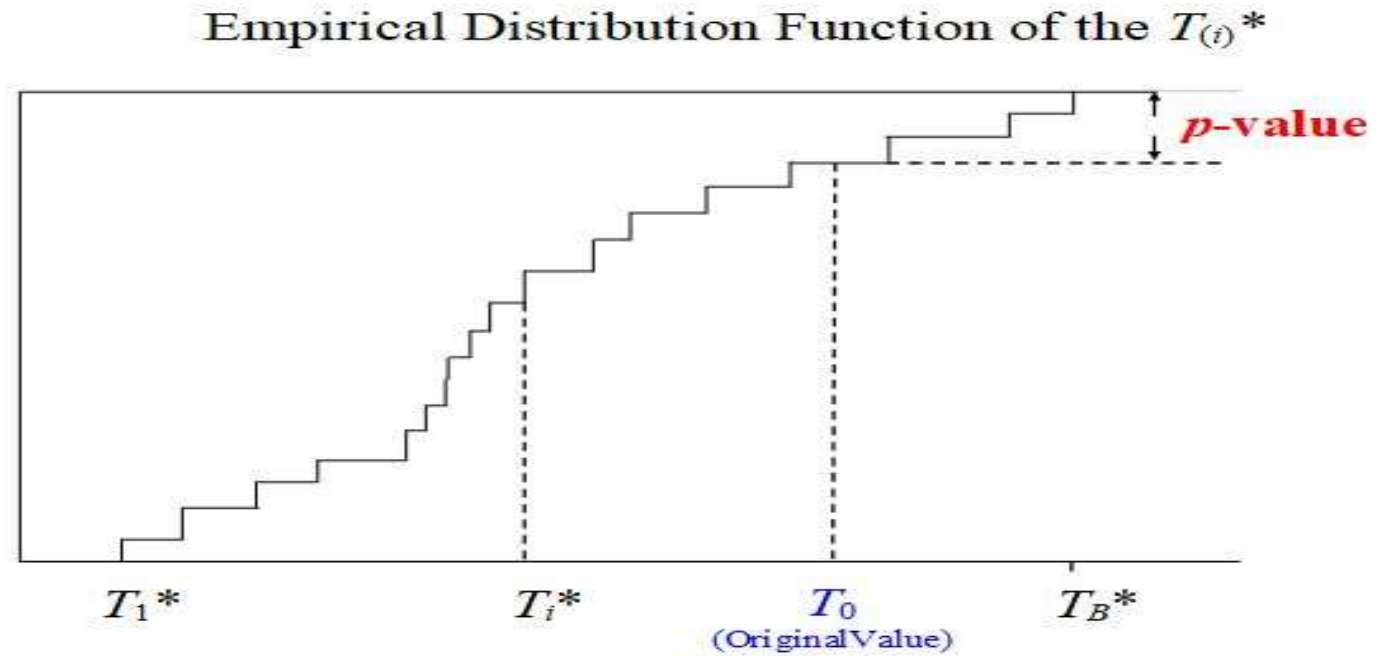
The EDF of the bootstrap sample estimates the distribution of *T*.



Often of interest is the *p*-value the original statistic value, T_0. This is

$$p = 1 - \tilde{F}^*(T_0).$$

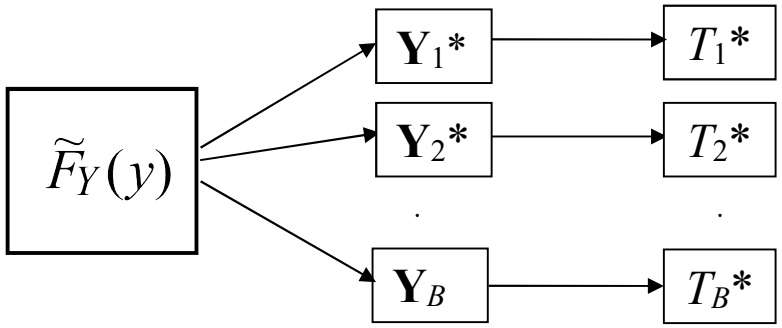
If the *p*-value is small then T_0 is in some sense unusual.



Excel Example 1 Here

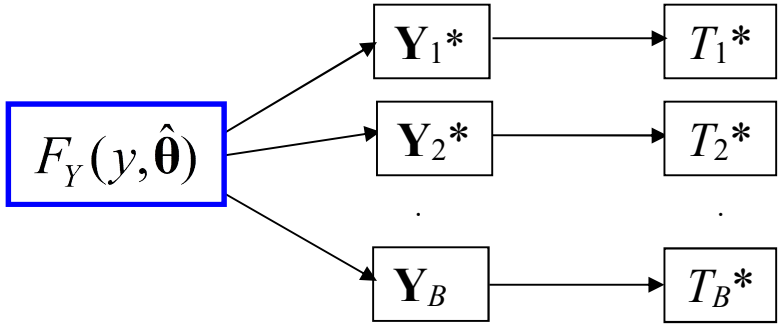
BasicBootstrapMeanMedian

Basic Bootstrap



If we have a parametric representation of $F_Y(y,\boldsymbol{\theta})$ with estimated parameters $\hat{\boldsymbol{\theta}}$

We can use the **Parametric Bootstrap**



Excel Example 2 Here ParametricBootstrap Mean/Median

We usually need to fit a parametric statistical model to data, as we did in our Example 2, need to use parametric bootstrapping. This has already been covered in the Course. I use a real data sample to remind you of what is needed.

The sample occurs in an Excel Toll Booth Example which we will also be using to discuss other issues in what follows.

The sample comprises 47 observed times in seconds taken to process vehicles at a toll booth waiting to cross a bridge.

4.3	10.9	4.7	4.7	3.1	5.2	6.7	4.5	3.6	7.2
6.6	5.8	6.3	4.7	8.2	6.2	4.2	4.1	3.3	4.6
6.3	4.0	3.1	3.5	7.8	5.0	5.7	5.8	6.4	5.2
8.0	10.5	4.9	6.1	8.0	7.7	4.3	12.5	7.9	3.9
4.0	4.4	6.7	3.8	6.4	7.2	4.8			

We suppose that these are gamma variates with PDF:

$$f_G(y,\alpha,\beta) = \Gamma^{-1}(\alpha)\beta^{-\alpha} y^{\alpha-1} \exp(-y/\beta)$$

We shall use *Maximum Likelihood Estimation*. to obtain

ML Estimates $\hat{\alpha}, \hat{\beta}$

Maximize the Log likelihood:

$L_G(\alpha,\beta,\mathbf{x}) = -n[\log(\Gamma(\alpha)) - \alpha\log\beta] - (\alpha-1)\sum_{j=1}^{n}\log(x_j) - \beta^{-1}\sum_{j=1}^{n} x_j$. A very convenient general numerical optimization method for doing this is the well-known simplex search procedure proposed by Nelder and Mead (1965).

$\boldsymbol{\theta} = (\alpha,\beta)$ ML estimator is $\hat{\boldsymbol{\theta}} = (\hat{\alpha},\hat{\beta})$

Asymptotic probability distribution of $\hat{\boldsymbol{\theta}}$ is known to be normal.

As the sample size $n \to \infty$,

$$\hat{\boldsymbol{\theta}} \sim N\{\boldsymbol{\theta}_0, \mathbf{V}(\boldsymbol{\theta}_0)\}$$

we can use

$$\hat{\boldsymbol{\theta}} \sim N\{\boldsymbol{\theta}_0, \mathbf{V}(\hat{\boldsymbol{\theta}})\}$$

where

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = [\, -\partial^2 \mathbf{L}(\boldsymbol{\theta}, \mathbf{y}) / \partial \boldsymbol{\theta}^2 \big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \,]^{-1}$$

The second derivative of the loglikelihood, $\partial^2 \mathbf{L}(\boldsymbol{\theta}, \mathbf{y}) / \partial \boldsymbol{\theta}^2$, that appears in the expression for $\mathbf{V}(\hat{\boldsymbol{\theta}})$ is called the *Hessian* (of $\mathbf{L}$). A numerical procedure is needed for this inversion.

A (1-α)100% confidence interval for the coefficient θ_1 is

$$\hat{\theta}_1 \pm z_{\alpha/2} \sqrt{V_{11}(\hat{\boldsymbol{\theta}})}$$

where $z_{\alpha/2}$ is the upper 100α/2 percentage point of the standard normal distribution.

Show Excel Examples 3 and 4 here

Example 3 gives the Gamma fit to Toll Booth Data. (show Optimize & Fit Sheets)

For comparison:

Example 4 gives the Normal fit to Toll Booth Data. (shoe Optimize & Fit Sheets)

Question: are either fits satisfactory?

Classical Goodness of Fit

Does the model that we have fitted actually fit the data very well?

Use a *goodness of fit test* (GOF test).
A popular test is the *chi-squared goodness of fit test*.

(i) The test statistic is easy to calculate
(ii) It has a *known* chi-squared distribution, under the null.

But (i) It is not all that powerful
(ii) The user has to decide how to group the data

The best GOF tests compare the fitted CDF $F_Y(y, \hat{\boldsymbol{\theta}})$ with the EDF $\widetilde{F}_Y(y)$

Such tests are called *EDF goodness of fit tests*.

The ***Anderson - Darling*** test, is the best by far. (Stephens, 1974)

But The critical values are very dependent on the model being tested

This means that different tables of test values are required for different models (see d'Agostino and Stephens, 1986).

***Anderson-Darling test statistic*:**

$$A^2 = \int \frac{\left(\widetilde{F}_Y(y) - F_Y(y)\right)^2}{F_Y(y)(1-F_Y(y))} dF_Y(y)$$

$$= -\sum_{i=1}^{n} (2i-1)[\ln Z_i + \ln(1 - Z_{n+1-i})]/n) - n$$

where $Z_i = F(Y_{(i)}, \hat{\boldsymbol{\theta}})$

The basic idea in using a goodness of fit test statistic is as follows:

If the sample has really been drawn from $F_0(y)$ then A^2 will not be large.

This follows from the Fundamental Theorem $\widetilde{F}(y) \to F_0(y)$

Thus A^2 will be a typical value. But what is a typical value?

Typical values given by its null distribution

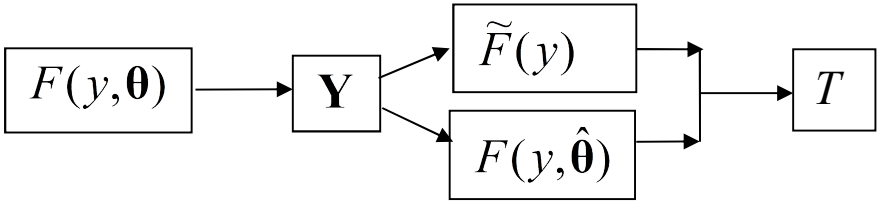
If the sample is drawn from a distribution different from $F_0(y)$ then A^2 will be large.

Its *p - value* will then be small.

This indicates that *T* has *not* been drawn from the supposed null distribution.

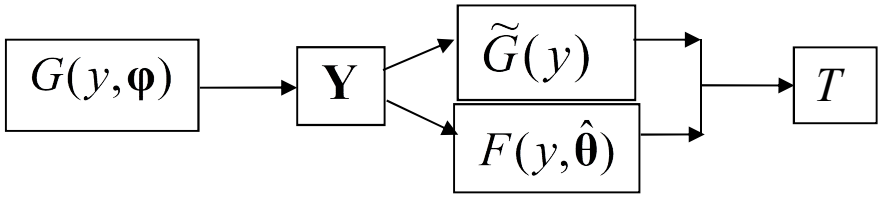
How a GOF Test Works

Null Case: Fitted model $F(y,\hat{\boldsymbol{\theta}})$ is the correct



Null Distribution where T is likely to be small

Alternative Case: Fitted model $F(y,\hat{\boldsymbol{\theta}})$ is an incorrect



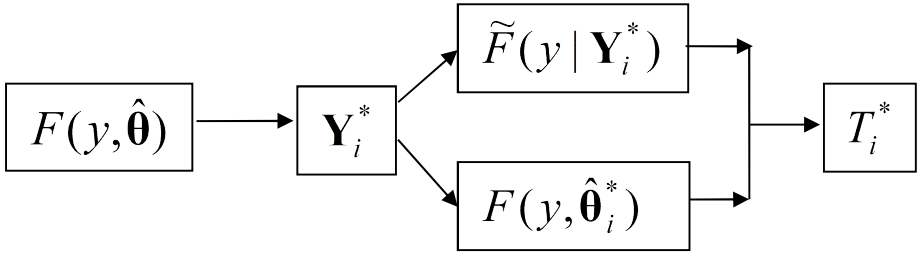
Non-Null Distribution where T is likely to be large

GOF test hinges on being able to calculate the null distribution of *T*.

The null distribution of the Anderson-Darling statistic is difficult to obtain. So not used as often as it should in practice.

Bootstrapping provides a simple and accurate way of resolving this problem.

Bootstrap Calculation of the Null Distribution of a GOF Test Statistic, *T*



$i = 1,2, \ldots, B$

Show Excel Examples 3 and 4 here again

GammaFitTollBooth Example only now including Bootstrapping Sheet
NormalFitTollBooth Example only now including Bootstrapping Sheet

Brief Summary of what we have considered so far

Our discussion has focused on how bootstrapping is useful for measuring the variability of a statistical quantity of interest.

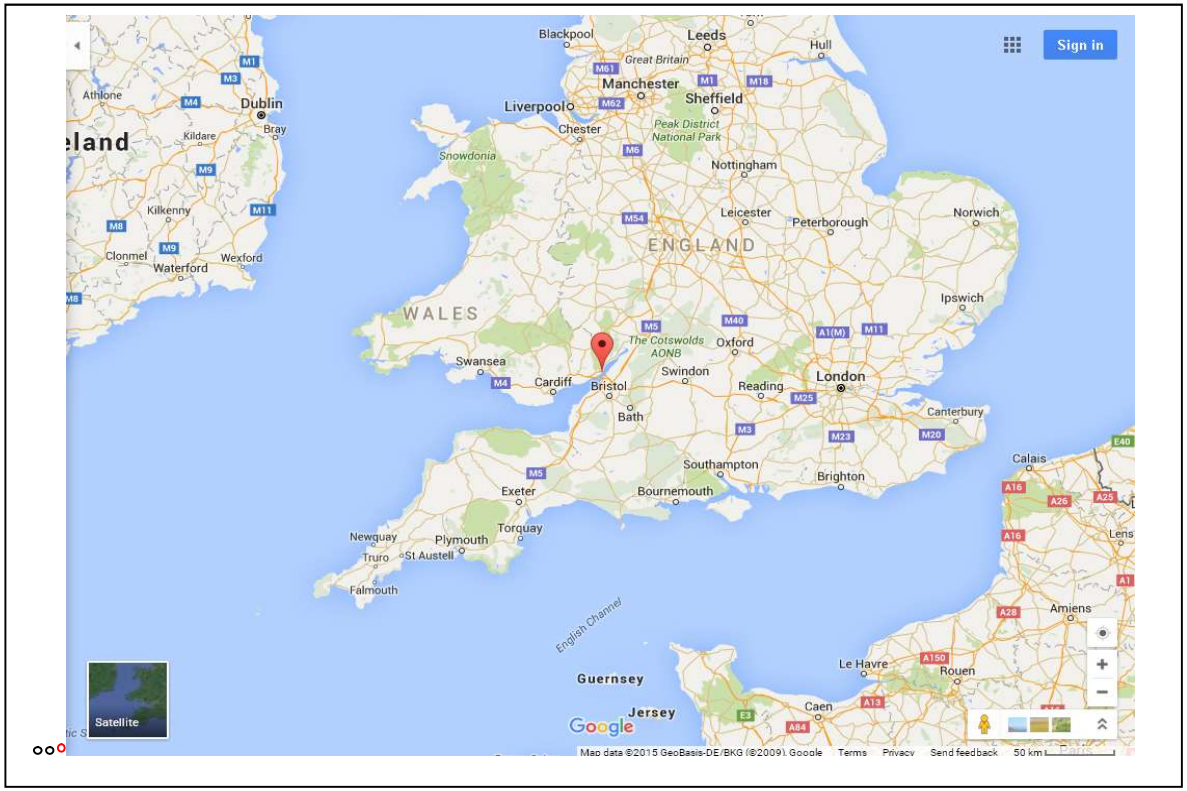
In the basic bootstrap, the bootstrap samples are drawn from the probability distribution of interest.

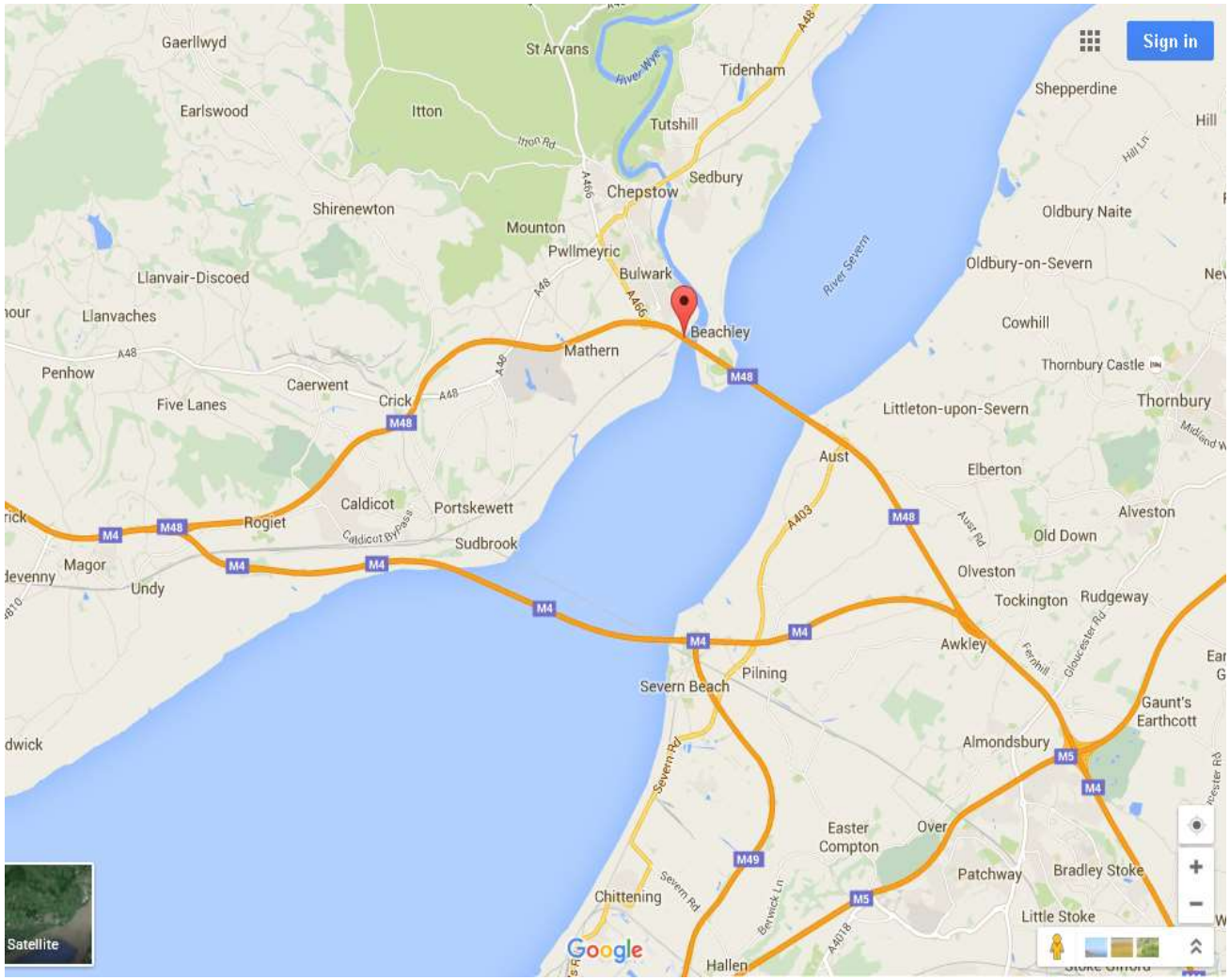
However when using parametric bootstrapping where we fit a parametric model which we have chosen, there is a problem if our selected distribution parametric distribution does not match the true distribution. However we have shown how use of GoF tests helps in selecting a suitable parametric distribution. In our Toll Booth Example we tried both the gamma and normal distributions.

The Toll Booth Example is a good example of **our next topic** which focuses on use of **Bootstrapping in Output Analysis**. First we describe this example in more detail.

Toll Booth Example

Operation of toll booths of the old Severn River bridge, UK, Griffiths and Williams (1984)




Sign in
Gaerllwyd
St Arvans
River Wye
Tidenham
Shepperdine
Earlswood
Itton
Tutshill
Hill
Chepstow
Sedbury
Shirenewton
Mounton
Pwllmeyric
Oldbury Naite
Oldbury-on-Severn
Llanvair-Discoed
Bulwark
River Severn
Llanvaches
Beachley
Cowhill
Mathern
Thornbury Castle
Penhow
Caerwent
Crick
Littleton-upon-Severn
Thornbury
Five Lanes
Aust
Elberton
Caldicot
Portskewett
Alveston
Rogiet
Old Down
Sudbrook
Olveston
Magor
Undy
Tockington
Rudgeway
Awkley
Pilning
Severn Beach
Gaunt's Earthcott
Almondsbury
Easter Compton
Over
Patchway
Bradley Stoke
Chittening
Little Stoke
Hallen
Google
Satellite

Unsatisfactory Original Bridge. Can you see why?

Each toll booth was modelled as a single server queue

Simulation model simulates the service of *l* vehicles.

Of interest: W(λ) - the average vehicle waiting time in the queue.

Service time data: Time taken for a vehicle to pay at the toll booth before crossing the bridge.

1 Parameter Uncertainty

Gamma service time parameters (Input Uncertainty)

Note that the arrival rate parameter not treated as part of parameter uncertainty as it is regarded as the argument of W(λ).

2 Simulation Uncertainty

Vehicle Waiting Time (When the functional form is not known and is numerically estimated by Simulation)

Use of Parametric Functions in Output Analysis

Suppose our real interest is not in the parameters themselves but in a function of $\boldsymbol{\theta}$, $g(\lambda, \boldsymbol{\theta})$, say, where $g(\lambda, \boldsymbol{\theta})$ is a function of λ, $\lambda_0 < \lambda < \lambda_1$

What is the MLE of $g(\lambda, \theta)$, $\lambda_0 < \lambda < \lambda_1$?

Answer is simple: The MLE of g is $\hat{g} = g(\lambda, \hat{\boldsymbol{\theta}})$.

Toll booth example: The steady state mean waiting time in the queue is known to be

$$W(\lambda|\alpha,\beta) = \frac{\lambda(1+\alpha)\alpha\beta^2}{2[1-\alpha\beta\lambda]} \quad \lambda_0 < \lambda < \lambda_1$$

Its ML estimated is simply $W(\lambda|\hat{\alpha},\hat{\beta})$ where we have replaced α, β by $\hat{\alpha},\hat{\beta}$:

An approximate (1-α)100% confidence interval for $g(\lambda, \boldsymbol{\theta})$ at a given λ is then

$$g(\lambda,\hat{\boldsymbol{\theta}}) \pm z_{\alpha/2}\sqrt{(\partial g/\partial\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}^{T}\,\mathbf{V}(\hat{\boldsymbol{\theta}})(\partial g/\partial\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}}$$

This is conventionally called **the delta-method**.

The above shows how to calculate Confidence Intervals for g(λ, θ)**, but these apply only for individual λ and are not suitable if confidence intervals for several diferent λ are needed simultaneously.**

Excel Examples 3 here again to give the Bootstrap Answer
Gamma Fit to Toll Booth Data now including the PerformanceIndex page

Note
I have used Performance Index (PI) and Performance Measure (PM) synonymously. In the case of the Toll Booth example the PI/PM is simply the expected waiting time W

Confidence Bands for Functions with Estimated Parameters

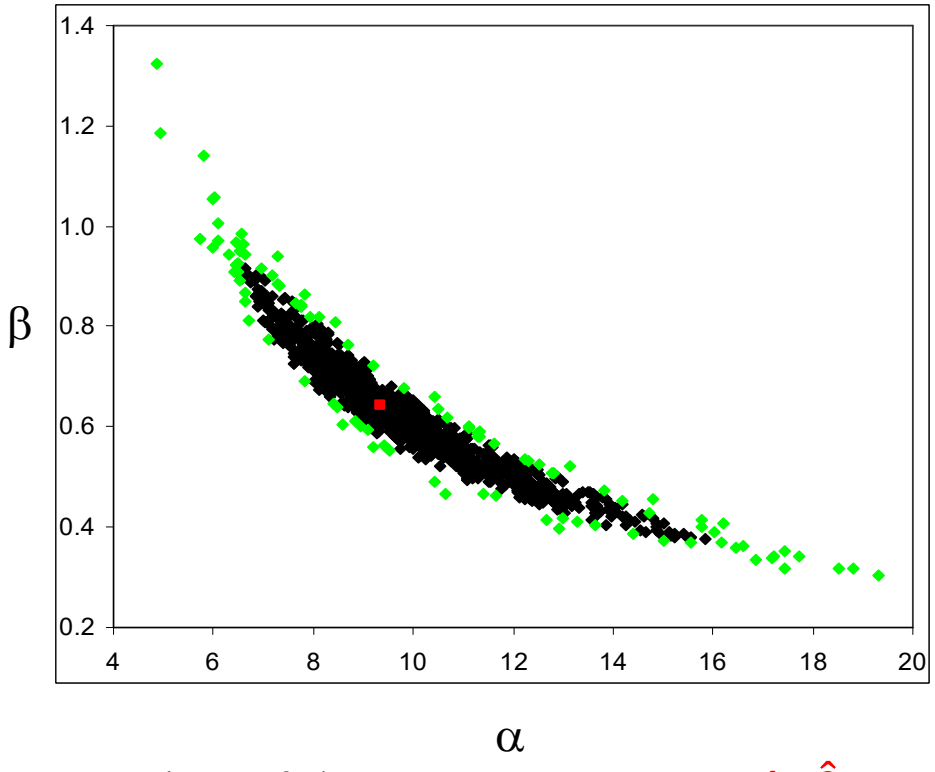
As we have aleady seen, a single **Confidence Interval** with given Confidence Level for the case $W(\lambda|\hat{\alpha},\hat{\beta})$ at a given λ.is straightforward. **But what about constructing a Band with upper and lower limits**?

$$\boldsymbol{WU}(\boldsymbol{\lambda}|\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) \quad \lambda_0 < \lambda < \lambda_1$$
$$\boldsymbol{WL}(\boldsymbol{\lambda}|\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) \quad \lambda_0 < \lambda < \lambda_1$$

Which will include the **entire** MLestimate $\boldsymbol{W}(\boldsymbol{\lambda}|\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) \quad \lambda_0 < \lambda < \lambda_1$ **and in which the true W lies with given Confidence Level?**

This question can be answered using Bootstrapping.

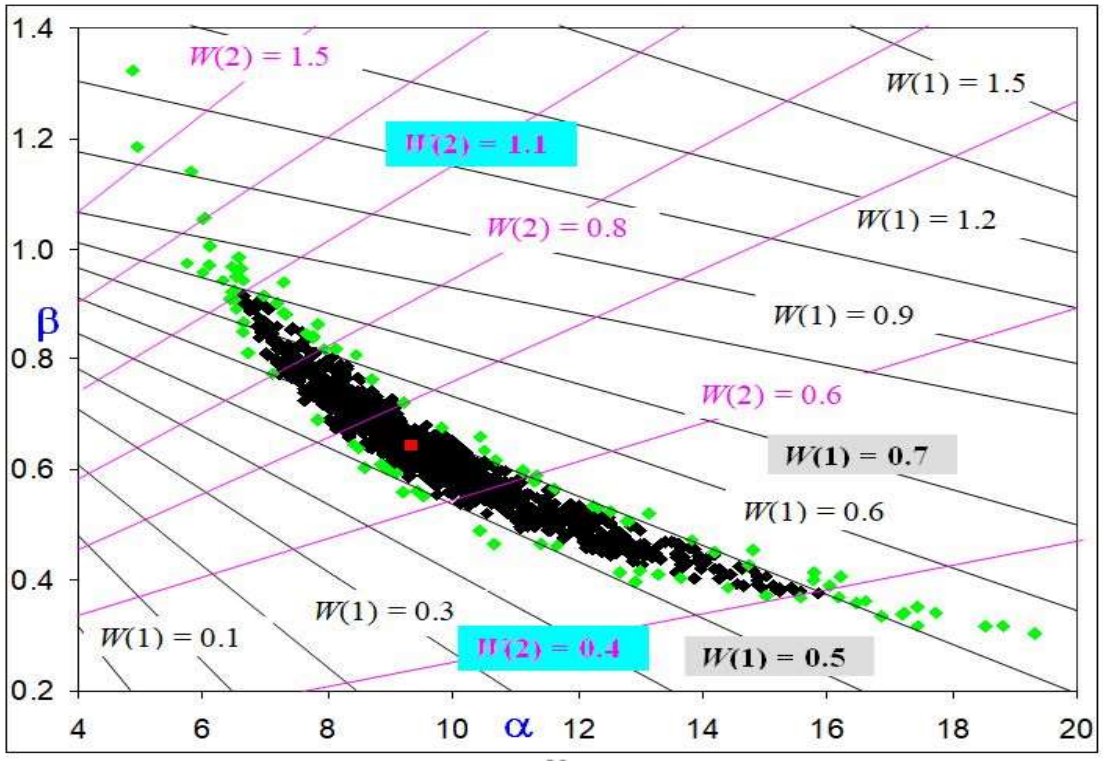
Use of ScatterPlot for calculating confidence bands



Red point: Location of the parameter MLEs $\hat{\alpha}, \hat{\beta}$

Black points:{ **R**} = 90% of the total number of points with highest likelihood values

Green points: {**Not in R**} = 10%, the Rest of the points with lowest likelihood values



Contours of $W = [\lambda(1+\alpha)\alpha\beta^2]/[2(1-\alpha\beta\lambda)]$ at $\lambda = 1$ and $\lambda = 2$

Confidence band is

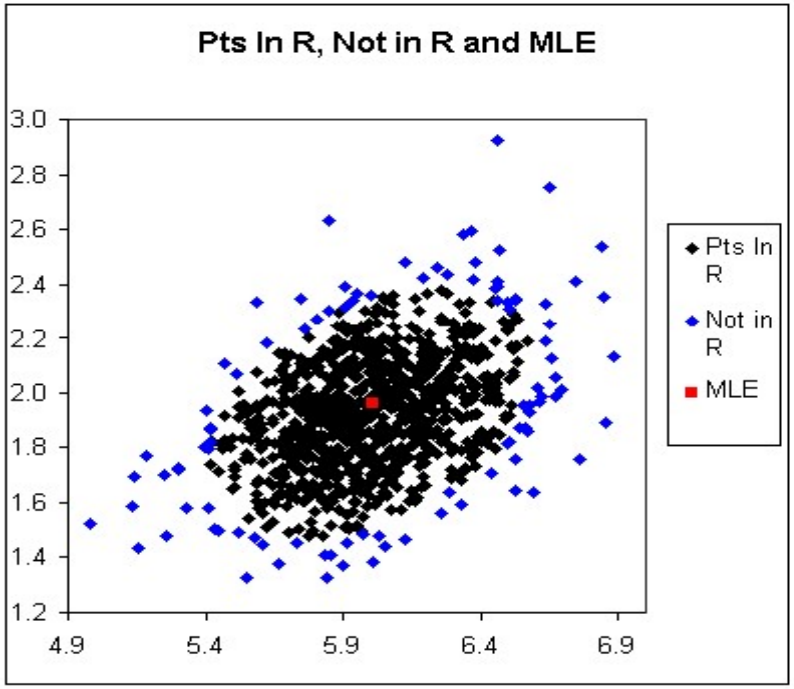
$W_{min}(\lambda) = \min_R W(\lambda|\, \alpha,\beta)$ $W_{max}(\lambda) = \max_R W(\lambda|\, \alpha,\beta)$ $\lambda_0 < \lambda < \lambda_1$

e.g.

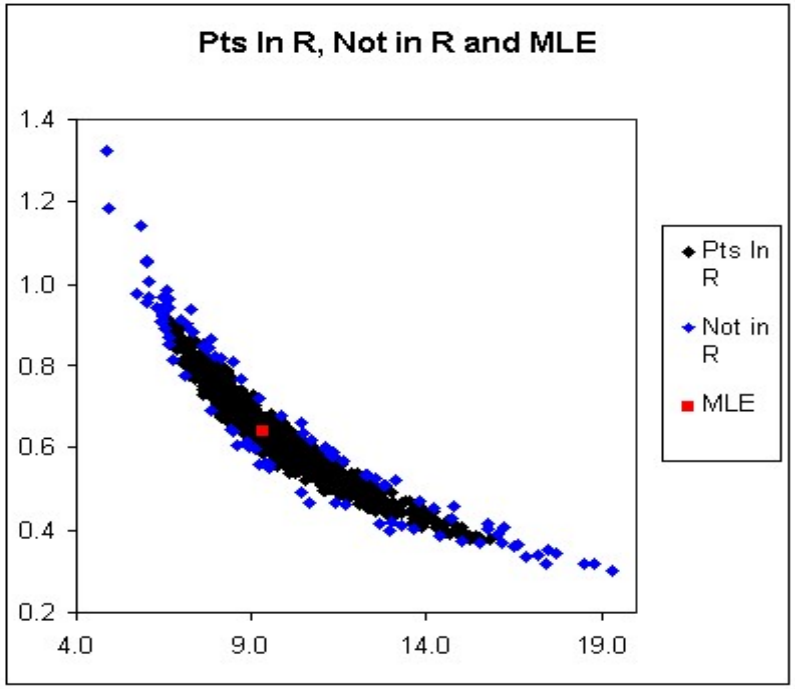
$W_{min}(1) = 0.5$ $W_{max}(1) = 0.7$ **and** $W_{min}(2) = 0.4$ $W_{max}(2) = 1.1$

Reparametrized parameters makes the band more accurate and symmetrical

$\mu = \alpha\beta \quad \sigma = \alpha\beta^2$

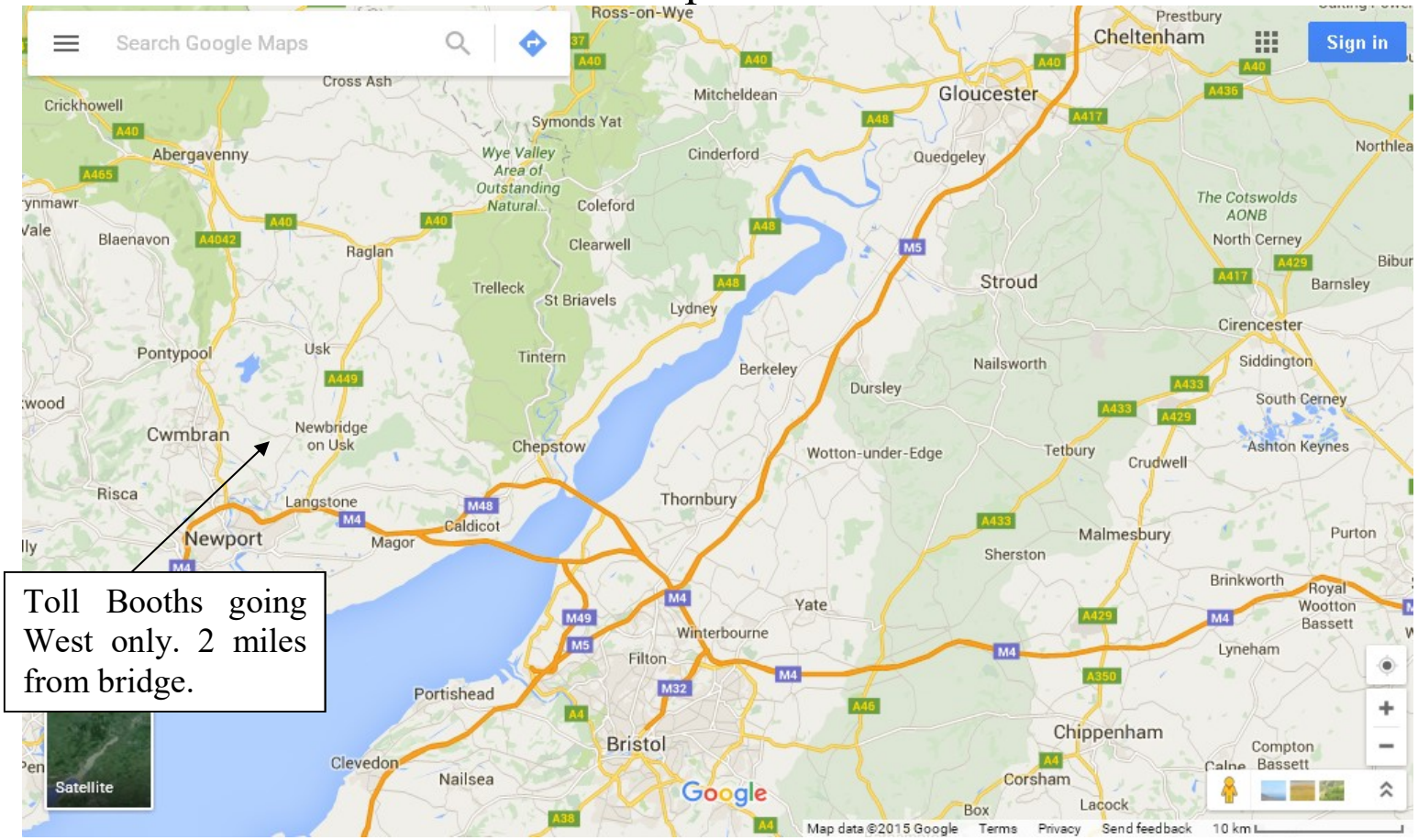


$\alpha = \mu^2/\sigma \quad \beta = \sigma/\mu$



Second Bridge. Built after OR Simulation Study

Additional Recommendation Adopted



Question: What happens when the Performance Measure is not a known mathematical function, but instead is a quantity that is obtained numerically from runs of the Simulation Model, so that it becomes an issue of Simulation Uncertainty?

The interesting answer is that, to first order of approximation, the overall variability of the PM when measured in terms statistical variance, is simply the sum of the variance of the Parameter Uncertainty and the variance of the Simulation Uncertainty. This was first pointed by Cheng and Holland (1997).

Using parametric bootstrapping, we can therefore do the following

Simply make B independent runs of the Simulation Model, where, in the ith run, the ith BS estimate, $\hat{\theta}^{(i)}$, of the vector of parameter is used. This allows us to estimate simultaneously the overall variability of the PM's obtained from these runs as in each run we are making independent runs so that there is simulation uncertainty, whilst the parameters will vary between runs so that there is parameter uncertainty.

Excel Example 5 Here

Final Summary of the uses of Parametric Bootstrapping

(1) It enables the accuracy of estimates of parameters to be assessed.

(2) It enables the suitability of fitted probability distributions to be assessed.

(3) It enables the accuracy of the estimate of Performance Measures (PM) to be assessed in terms of Parameter Uncertainty and Simulation Uncertainty as defined by Cheng and Holland (1997); whether the mathematical form of the PM is a known function of the parameters or not.

I have not had the time to include my simulation modelling work on Covid-19 with three other colleagues, two of whom worked for World Health Organisation. I introduced them to Professor Currie, then a research student of mine, worked with them. I have put on my Southampton Personal Home page **Power Point Slides** of a presentation I gave at SW20 (postponed from last year!), the March OR Simulation Workshop. It is on modelling PreSymptomatic (now called Asymptomatic) Infectiousness. **Do have a look as much of it is still very relevant, especially concerning herd-immunity.**

References

Cheng, R C H and Holland, W. (1997). Sensitivity of Computer Simulation Erros in Input Data. *J. Statist. Comput. Simul.*, **57**, 219-241.

Cheng, R C H (2017) *Non-Standard Parametric Statistical Inference*, Oxford University Press.